V. PROJECT DESCRIPTIONS

CORE RESEARCH & DEVELOPMENT

A. DENDRAL

Project: DENDRAL
Realtime

Investigator: Edward Feigenbaum,
Joshua Lederberg, and Carl Djerassi

Dept. of Chemistry, Computer Science,
and Genetics

The DENDRAL project involves collaboration between the Instrumentation Research Laboratory operating under NASA grant NGR-05-020-004, investigators operating under NIH grant RR00612, and ACME.

The emphasis of the DENDRAL-ACME efforts is computer science, while that of IRL-ACME endeavors is data acquisition and computer instrument control.

The DENDRAL project aims at emulating in a computer program the inductive behavior of the scientist in an important but sharply limited area of science; organic chemistry. Most of the work is addressed to the following problem; given analytic data (the mass spectrum) of an unknown compound, infer a workable number of plausible solutions, that is, a small list of candidate molecular structures. In order to complete the task, the DENDRAL program then deduces the mass spectrum predicted by the theory of mass spectrometry for each of the candidates and selects the most productive hypothesis, i.e., the structure whose predicted spectrum must closely matches the data.

The project has designed, engineered, and demonstrated a computer program that manifests many aspects of human problem solving techniques. It also works faster than human intelligence in solving problems chosen from an appropriately limited domain of types of compounds, as illustrated in the cited publications.

Some of the essential features of the DENDRAL program include:

Conceptualizing organic chemistry in terms of topological graph theory, i.e., a general theory of ways of combining atoms.

Embodying this approach in an exhaustive HYPOTHESIS GENERATOR. This is a program which is capable, in principle, of "imagining" every conceivable molecular structure.

Organizing the GENERATOR so that it avoids duplication and irrelevancy, and moves from structure to structure in an orderly and predictable way.

Core Research & Development (Continued)

The key concept is that induction becomes a process of efficient selection from the domain of all possible structures. Heuristic search and evaluation are used to implement this "efficient selection."

Most of the ingenuity in the program is devoted to heuristic modifications of the GENERATOR. Some of these modifications result in early pruning of unproductive or implausible branches of the search tree. Other modifications require that the program consult the data for cues (pattern analysis) that can be used by the GENERATOR as a plan for a more effective order of priorities during hypothesis generation. The program incorporates a memory of solved sub-problems that can be consulted to look up a result rather than compute it over and over again. The program is aimed at facilitating the entry of new ideas by the chemist when discrepancies are perceived between the actual functioning of the program and his expectation of it.

The DENDRAL research effort has continued to develop along several dimensions during Fiscal 1973. The mass spectra of some previously uninvestigated compounds were recorded. The computer program has been extended to analyze the mass spectra of a more complex class of compounds, using new kinds of data. The artificial intelligence work on theory formation and program generality has also progressed.

The techniques of artificial intelligence have been applied successfully for the first time to a problem of direct biological relevance, namely the analysis of the high resolution mass spectra of estrogenic steriods. The performance of this program has been shown to compare favorably with the performance of trained mass spectroscopists. (see Smith, et al. (1972)

Of particular significance in this effort were, in addition to exceptional performance, the potential for analysis of estrogens <u>without prior separation</u>, and for generalization of the programming approach to other classes of molecules.

Because of the structure of the Heuristic DENDRAL program for estrogens, it is immaterial whether the spectrum to be analyzed is derived from a single compound or a mixture of compounds. Each component is analyzed, in terms of molecular structure, in turn, independently of the other components. This facility, if successful in practice, would represent a significant advance of the technique of mass spectrometry. Many problem areas, because of physical characteristics of samples or limited sample quantities, could be successfully approached utilizing the spectra of the unseparated mixtures. Even in combined gas chromatography/mass spectrometry (GC/MS), many mixture components will be unresolved and an analysis program must be capable of dealing with these mixtures.

Core Research & Development (Continued)

We have, in collaboration with Prof. H. Adlercreutz of the University of Helsinki, recently completed a series of analyses of various fractions of estrogens extracted from body fluids and supplied to us by Prof. Adlercreutz. These fractions (analyzed by us as unknowns) were found to contain between one and four major components, and structural analysis of each major component was carried out successfully by the above program. These mixtures were analyzed as unseparated, underivatized compounds. The implications of this success are considerable. Many compounds isolated from body fluids are present in very small amounts and complete separation of the compounds of interest from the many hundreds of other compounds is difficult, time-consuming and prone to result in sample loss and contamination. We have found in this study that mixtures of limited complexity, which are difficult to analyze by conventional GC/MS techniques without derivatization (which frequently makes structural analysis more difficult), can be rationalized even in the presence of significant amounts of impurities. A manuscript on this study has been submitted to the Journal of the American Chemical Society.

In the past year we have extended our library of high resolution mass spectra of estrogens to include 67 compounds. These data represent an important resource and have been included (as low resolution spectra for the moment) in a collection of mass spectra of biologically important molecules being organized by Prof. S. Markey at the University of Colorado.

The Heuristic DENDRAL program for complex molecules has received considerable attention during the last year in order to remove compound class specific information or program strategies. By removing information which is specific to estrogens, the program has become much more general. This effort has resulted in a production version of the program which is designed to allow the chemist to apply the program to the analysis of the high resolution mass spectrum of any molecule with a minimum of effort. Given the spectrum of a known or unknown compound, the chemist can supply the following kinds of information to guide analysis of the mass spectrum: a) Specifications of basic structure (superatom) common to the class of molecules. b) Specification of the Fragmentation rules to be applied to the superatom, in the form of bond cleavages, hydrogen transfers and charge placement. c) Special rules on the relative importance of the various fragments resulting from the above fragmentations. d) Threshold settings to prevent consideration of low intensity ions. e) Available metastable ion data and the way these data are subsequently used -- to establish definitive relationships between fragment ions and their respective molecular ions. f) Available low ionizing voltage data -- to aid the search for molecular ions. g) Results of deuterium exchange of labile hydrogens -- to specify the number of, e.g., -OH groups.

Core Research & Development (Continued)

We have been very successful in testing the generality of the program, with particular emphasis on other classes of biologically important molecules. We have used the program in analysis of high resolution mass spectra of progesterone and some methylated analogs, a small number of androstane/testosterone related compounds, steroidal sapogenins and n-butyl-trifluoroacetyl derivatives of amino acids.

The Heuristic DENDRAL performance program described above is an automated hypothesis formation program which models "routine", day-to-day work in science. In particular, it models the inferential procedures of scientists identifying components, such as those found in human body fluids. The power of this program clearly lies in its knowledge about various classes of compounds normally found in body fluids, which knowledge allows identification of the compounds.

The Meta-DENDRAL program described in this part is a critical adjunct to the performance program because it is designed to supply the knowledge which the performance program uses. Theory formation is essential in order to carry out the routine analyses - either by hand or by computer. However, the staggering amount of effort required to build a working theory (even for a single class of compounds) holds back the routine analyses. The goal of the Meta-DENDRAL program is to form working theories automatically (from collections of experimental data) and thus reduce the human effort required at this stage. By speeding up the time between collecting data for a class of compounds and understanding the rules underlying the data, the Meta-DENDRAL program will thus provide an improvement in the development of diagnostic procedures.

Detailed accounts of this research are available in the DENDRAL Project annual report to the National Institutes of Health, in several papers already published, and in manuscripts submitted for publication.

1. For pertinent reviews see: C. G. Hammar, B. Holmstedt, J. E. Lindgren and R. Tham, Advan. Pharma.Col. Chemother., 7, 53, (1969); J. A. Vollmin and M. Muller, Enzymol. Biol. Clin., 10, 458 (1969).

2. J. R. Althans, K. Biemann, J. Biller, P. F. Donaghue, D. A. Evans, H. J. Forster, H. S. Hertz, C. E. Hignite, R. C. Murphy, G. Petrie and V.Reinhold, Experientia, 26, 714 (1970).

3. H. Fales, G. Milne and N. Law, reported in Medical World News, February 19, 1971.

4. E. Jellum, O. Stokke and L. Eldjarn, The Scandinavian Journal of Clinical and Laboratory Investigation, 27, 273 (1971).

5. A. L. Burlingame and G. A. Johanson, Anal. Chem., 44, 337R (1972).

Core Research & Development (Continued)

6. H. S. Hertz, R. A. Hites and K. Biemann, Analytical Chemistry, 43, 681 (1971), S. L. Grotch, ibid., 43, 1362 (1971).

7. E. A. Feigenbaum, B. G. Buchanan, and J. Lederberg, "On Generality and Problem Solving: A Case Study Using the DENDRAL Program". In Machine Intelligence 6 (B. Meltzer and D. Michie, eds.) Edinburgh University Press (1971). (Also Stanford Artificial Intelligence Project Memo No. 131.)

8. A. Buchs, A. B. Delfino, C. Djerassi, A. M. Duffield, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, G. Schroll, and G. L. Sutherland, "The Application of Artificial Intelligence in the Interpretation of Low-Resolution Mass Spectra", Advances in Mass Spectrometry, 5, 314.

9. B. G. Buchanan and J. Lederberg, "The Heuristic DENDRAL Program for Explaining Empirical Data". In proceedings of the IFIP Congress 71, Ljubljana, Yugoslavia (1971). (Also Stanford Artificial Intelligence Project Memo No. 141.)

10. B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg, "A Heuristic Programming Study of Theory Formation in Science." In proceedings of the Second International Joint Conference on Artificial Intelligence, Imperial College, London (September, 1971). (Also Stanford Artificial Intelligence Project Memo No. 145.)

11. Buchanan, B. G.,Duffield, A. M.,Robertson, A. V., "An Application of Artificial Intelligence to the Interpretation of Mass Spectra", Mass Spectrometry Techniques and Appliances, Edited by George W. A. Milne, John Wiley & Sons, Inc., 1971, p. 121-77.

12. D. H. Smith, B. G. Buchanan, R. S. Engelmore, A. M. Duffield, A. Yeo, E. A. Feigenbaum, J. Lederberg, and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference VIII. An approach to the Computer Interpretation of the High Resolution Mass Spectra of Complex Molecules. Structure Elucidation of Estrogenic Steroids", Journal of the American Chemical Society, 94, 5962-5971 (1972).

13. B. G. Buchanan, E. A. Feigenbaum, and N. S. Sridharan, "Heuristic Theory Formation: Data Interpretation and Rule Formation". In Machine Intelligence 7, Edinburgh University Press (1972).

14. Brown, H., Masinter L., Hjelmeland, L., "Constructive Graph Labeling Using Double Cosets". Discrete Mathematics (in press), (Also Computer Science Memo 318, 1972.)

15. B. G. Buchanan, Review of Hubert Dreyfus' "What Computers Can't Do: A Critique of Artificial Reason", Computing Reviews (January, 1973). (Also Stanford Artificial Intelligence Project Memo No. 181)

Core Research & Development (Continued)

16. D. H. Smith, B. G. Buchanan, R. S. Engelmore, H. Aldercreutz and C. Djerassi, "Applications of Artificial Intelligence for Chemical Inference IX. Analysis of Mixtures Without Prior Separation as Illustrated for Estrogens". Submitted to the Journal of the American Chemical Society.

17. D. H. Smith, B. G. Buchanan, W. C. White, E. A. Feigenbaum, C. Djerassi and J. Lederberg, "Applications of Artificial Intelligence for Chemical Inference X. Intsum. A Data Interpretation Program as Applied to the Collected Mass Spectra of Estrogenic Steroids". To be submitted.

The preceding comments on DENDRAL involve Parts A and C as described in the table below. The balance of this section deals with Part B, instrumentation aspects.

Part A: Applications of Artificial Intelligence to Mass Spectrometry.

Part B(i): Mass Spectrometer Data System Development.

Part B(ii): Analysis of the Chemical Constituents of Body Fluids.

Part C: Extending the Theory of Mass Spectrometry by Computer.

ACME computer support for DENDRAL Part B has been treated as ACME core research activity during FY73. Excerpts from DENDRAL's annual report follow, detailing recent accomplishments.

The large volume of data which must be reduced and interpreted from each GC/MS analysis of a body fluid sample together with the increasing number of samples which must be processed to be responsive to clinical needs, point to more and more highly automated and reliable GC/MS systems. This portion of the proposal addresses the problems of developing and applying such automated systems from several points of view. First, we propose to investigate the integration of sophisticated computer analysis programs into data reduction, data interpretation, and instrument management functions in order to progressively relieve the chemist from manually performing these tasks. Second, we will maintain the daily operation of our GC/MS systems for the on-going investigation of clinical applications and the acquisition of data necessary for the development of automated interpretation programs.

Our overall objectives for automating GC/MS systems comprise a number of specific subgoals including a) implementing highly automated and reliable systems for the acquisition and reduction of low resolution, high resolution, and metastable mass spectral data; b) implementing a data system to support combined gas chromatography/high resolution mass spectrometry; c) automating the location and identification of constituents of body fluid extracts from gas chromatogram and mass spectrum information for the routine application of these techniques to clinical problems; and d) investigating the intelligent closed loop control of mass spectrometer systems in order to optimize the data acquired relative to the task of data interpretation.

Core Research & Development (Continued)

A. Mass Spectrometer Data System Automation

Concentrating initially on the MAT-711 spectrometer, we have made significant progress toward a reliable, automated data acquisition and reduction system for scanned low and high resolution spectra. This system is largely failsafe and requires no operator support or intervention in the calculation procedures. Output and warnings to the operator are provided on a CRT adjacent to the maas spectrometer. The system contains many interactive features which permit the operator to examine selected features of the data at his leisure. The feedback currently provided to the operator to assist in instrument set-up and operation can just as well be routed to hardware control elements for these functions thereby allowing computer maintenance of optimum instrument performance.

Progress in this area is an integration of our efforts in hardware and software improvements:

HARDWARE - The basic system consists of the mass spectrometer interfaced to a PDP-11/20 computer for data acquisition, pre-filtering, and time buffering into the ACME time-shared 360/50. The more complex aspects of data reduction are done in the 360/50 since the PDP-11 has limited memory and arithmetic capabilities. New interfaces for mass spectrometer operation and control have been developed. The interfaces can handle (through an analog multiplexer) several analog inputs and outputs which require that the PDP-11 computer be relatively near the mass spectrometer. We now have the capability for the following kinds of operation through the new interfaces.

i) Computer selection of digitization rate.

ii) Computer selection of data path (interrupt mode or direct memory access (DMA).

iii) Direct memory access for faster operation in the data acquisition mode.

iv) Computer selection of analog input and output channels.

v) Sensing of several analog channels through a multiplexer (e.g., ion signal, total ion current).

vi) Magnet scan control. This control can be exercised manually or set by the computer. It controls both time of scan and flyback time. Coupled with selection of scan rate, any desired mass range can be scanned at any desired scan rate.

vii) The computer can monitor the mass spectrometer's mass marker output as additional information which will be used to effect calibration.

Core Research & Development (Continued)

SOFTWARE - Automatic instrument calibration and data reduction programs have been developed to a high degree of sophistication. We can now accurately model the behavior of the MAT-711 mass spectrometer over a variety of scan rates and resolving powers. Our instrument diagnostic routines are depended upon by the spectrometer operator to indicate successful operation or to help point to instrument malfunctions or set-up errors. Some features of these programs are described below.

i) Data Acquisition. Programs have been written which permit acquisition of peak profile data at high data rates using the PDP-11 as an intermediate data filter and buffer store between the mass spectrometer and ACME. This allows data acquisition to proceed even under the time constraints of the time-sharing system. Storage of peak profiles rather than all data collected has greatly reduced the storage requirements of the program and saves time as the background data (below threshold) are removed in realtime. An automatic thresholding program is in operation which statistically evaluates background noise and thresholds subsequent data accordingly. Amplifier drift can thus be compensated. We have developed some theoretical models of the data acquisition process which suggest that high data acquisition rates are not necessary to maintain the integrity of the data. Demonstration of this fact with actual data has helped relieve the burden of high data rates on the computer system, particularly as imposed by GC/MS operation, and permits more data reduction to be accomplished in realtime or alternatively reduces the required data acquisition computer capacity.

ii) Instrument Evaluation. A high resolution mass spectrometer operating in a dynamic scanning mode is a complex instrument and many things can go wrong which are difficult for the operator to detect in realtime. In order for the computer to assist in maintaining data quality, it must have a model of spectrometer operation on the basis of which data quality can be assessed and processing suitably adapted as well as instrument performance optimized. We have developed a program which monitors the state of the mass spectrometer.

iii) Data Reduction. A program has been written which allows automatic reduction of high resolution data based on the results of the prior instrument evaluation data. Conversion of peak positions in time to the corresponding mass values is effected by mapping each spectrum into the calibration model developed previously. The interpolation algorithm between reference calibration points incorporates

Core Research & Development (Continued)

a quadratically varying exponential time constant to account for the second order character of a magnet discharging through a resistance and a capacitance as well as an offset at infinite time to account for residual magnetization affecting accuracy at low masses.

Perfluorokerosene (PFK) peaks, introduced into high resolution mass spectra for internal mass calibration, are distinguished from unknown peaks by a pattern recognition algorithm which compares the relationships between sequences of reference peaks in the calibration run with the set of possible corresponding sequences in the sample run. The candidate sequence is selected which best approximates calibrated performance within constraints of internally consistent scan model variations. This approach minimizes the need for selection criteria such as greatest negative mass defect for reference peaks, the validity of which cannot be guaranteed. Excellent performance results from using sequences containing 10 reference peaks.

Unresolved peaks are separated by a new analytical algorithm, the operation of which is based on a calculated model peak derived from known singlet peaks rather than the assumption of a particular parametric shape (e.g., triangular, Gaussian, etc.) This alogorithm provides an effective increase in system resolution by a factor of three thereby effectively increasing system sensitivity. By measuring and comparing successive moments of the sample and model peaks, a series of hypotheses are tested to establish the multiplicity of the peak, minimizing computing requirements for the usually encountered simple peaks. Analytic expressions for the amplitudes and positions of component peaks have been derived in the doublet case in terms of the first four moments of the peak complex. This eliminates time consuming iteration procedures for this important multiplet case. Iteration is still required for more complex multiplets.

Elemental compositions are calculated from high resolution mass values with a new, efficient table look-up algorithm developed by Lederberg.

Future work will extend these ideas to a system for the acquisition of selected metastable information as well as to include the quadrupole system used in the routine low resolution clinical work.

B. GAS Chromatography/High Resolution Mass Spectrometry.

We have recently verified the feasibility of combined gas chromatography/high resolution mass spectrometry (GC/HRMS). Using the programs described above we can acquire selected scans and reduce them automatically,

Core Research & Development (Continued)

although the procedures are slow compared to "realtime" due to the limitations of the time-shared ACME facility. We have recorded sufficient spectra of standard compounds to show that the system is performing well.

We have begun to exercise the GC/HRMS system on urine fractions containing significant components whose structures have not been elucidated on the basis of low resolution spectra alone. Whereas more work is required to establish system performance capabilities, two things have become clear: 1) GC/HRMS will be a useful analytical adjunct to our low resolution GC/MS clinical studies to assist in the identification of significant components whose structures are not elucidated on the basis of low resolution spectra alone, and 2) the sensitivity of the present system limits analysis to relatively intense GC peaks.

Recent experiments in operation of the mass spectrometer in conjunction with the gas chromatograph have also shown that the present ACME computer facility cannot provide the rapid service required to acquire repetitive scans at either high or low resolving powers. We can, however, acquire scans on a periodic basis, meaning most GC peaks in a run can be scanned once at high resolving power. We are presently implementing a disk on the PDP-11 to act as a temporary data buffer between the mass spectrometer and ACME. This disk will allow acquisition of repetitive scans, while data reduction must be deferred to completion of the GC run.

C. Automated GC/MS Data Reduction

The application of GC/MS techniques to clinical problems as described in Part B(ii) of this proposal has made clear the need for automating the analysis of the results of a GC/MS experiment. Previous paragraphs dealt with the problems of reducing raw data in preparation for analysis. At this point the data must be analyzed with a minimum of human interaction in terms of locating and identifying specific constituents of the GC effluent. The problem of identification is addressed by the library search and DENDRAL mass spectrum interpretation programs discussed in Part A of this proposal. The problem of locating effluent components in the GC/MS output involves extracting from the approximately 700 spectra collected during a GC run, the 50 or so representing components of the body fluid sample. The raw spectra are in part contaminated with background "column bleed" and in part composited with adjacent constituent spectra unresolved by the GC.

We have begun to develop a solution to this problem with very promising results.

Core Research & Development

D. Closed-Loop Instrument Control.

The task of collection of different types of mass spectral information (e.g., high resolution spectra, low ionizing voltage spectra snd selected metastable information) under closed loop control during a GC/MS experiment is extremely difficult and may not be realizable with current technology. We are studying this problem in a manner which will allow the system to be used for important research problems (e.g., routine analysis of urine fractions without fully closed loop control) while aspects of instrument control strategy are developed in an incremental fashion.

Core Research & Development (Continued)

B. Time Oriented Database System (TOD)

Investigator: Dr. James Fries
and the ACME Staff

Project: J_FRIES_.DATABANK
F_GERMAN.TOD
DATABANK.TODD

Dept. of Medicine - Immunology
and ACME

In 1970 and 1971, Dr. James Fries in the Division of Immunology of the Department of Medicine developed concepts and implemented programs which he labeled "Time Oriented Database". One of the first steps was the development of standard forms for use in the medical record. These forms are completed manually and require no computer intervention or interaction. Use of the new medical record forms has proved highly desirable in several clinics at Stanford since that time, with or without the associated computer programs. The relationship of the computer to the project makes possible rapid comparison and statistical analysis of various data items covering multiple visits for one patient or for many patients.

In the summer of 1972, a design study was completed which would generalize the use of the TOD programs on the ACME system so that several divisions could use a common set of programs. The design effort was handled primarily by Stephen Weyl with assistance from Gio Wiederhold and Frank Germano. Implementation of the new generalized TOD programs was managed by Frank Germano with Stephen Weyl, Rick Giusti, Bob Bassett, and Jane Whitner handling the programming.

As of May 1, 1973, several TOD databanks had been implemented and several more had been planned. The table below reflects the progress to that date.

TOD Implementation Progress Report (May 1, 1973)

PRESENT TOD DATABANKS

User	Medical Speciality	Comments
Dr. Jim Fries	Immunology	Operational on TOD 3 months
Dr. S. Rosenberg Dr. L. William	Oncology	Operational on TOD 3 months
Dr. M. Stern	Metabolic Disease Clinic	Databank defined. Time-Oriented Medical Record forms being printed. Data entry will begin when forms are ready.

TOD DATABANKS GOING THROUGH DEFINITION PROCESS

Dr. K. Brodie Dr. M. Rosenzweig	Psychiatry Alcohol & Violence Prevention Clinic	
F. Germano	TOD Group	TOD group mailing list
D. Lombardi	Student Affairs Office	Part of the TOD System will be used to set up a Medical Student Record System
Dr. Bleck	Childrens Hospital Orthopedic Service	TOMR forms designed. Waiting to define databank.
Dr. J. Gamel	Ophthalmology Clinic	Databank defined. Presently collecting input data.

GROUPS CONSIDERING A TOD DATABANK DEFINITION

Dr. Wilbur	Childrens Hospital
Dr. Miller	Childrens Hospital
Dr. V. Johnson Dr. A. Hackel	Pediatrics
Dr. M. Bagshaw	Radiology

The system which was announced in January 1973 is but a first step in development of database systems at Stanford. Clearly more development effort will follow which will improve the data entry techniques to be employed, enhance quality control of data entered, and increase the amount of shared data in the files.

The following pages contain four ACME Notes written to document the TOD system, along with explanatory remarks. The four ACME Notes are:

TODI - Introduction to TOD System
TODREF - Index to TOD ACME Notes
TODDDL - TOD Databank Description Language
TODCST - Analyzing the Costs of Running a TOD Databank

ACME Note

TODI-3

Frank Germano/Steve Weyl

Introduction to TOD (Time-Oriented Databank) System

April 5, 1973

The TOD system is a set of programs available from ACME designed to aid users in the creation, maintenance, and use of computer "databanks" which store patient-related information over time. These programs are available as TOD public programs on the PL/ACME system. If a user can conceptually view his patient data in the form of a three-dimensional array, indexed by patient, parameter, and time, he can use the TOD system. A recently conducted database review of medical data stored on the ACME computer system and other Stanford computers revealed that many of these databanks have this form. The results of this survey are summarized in ACME Note DBS.

Flexibility and Independence

In order to offer a system of programs which support most patient-related databases implemented on the ACME facility, a large degree of flexibility and independence had to be built into the system. The TOD approach is a decentralized one, in which each division maintains a separate databank, whose inter-relation to all databanks is well defined. Each TOD databank is set up and used under one ACME name and project. The databank planner is the administrator of that databank, not ACME. To provide for user definition, an extra file is added to the databank. This file is called a SCHEMA file; it describes the form of the databank. It stores that information which makes each TOD databank unique for its user. Public programs which act on a TOD databank look to this file for descriptive information about the databank. This information is then used by the various programs which act on the databank during their operation.

Advantages of TOD for A User

Use of the TOD system can offer several advantages to the user. Some of the direct advantages are discussed below.

1. Less Effort to Utilize a Patient-Related Databank:

Prior to TOD each user essentially had to write programs to set up, maintain, and use a databank. This represented a great duplication of effort. The databanks tended to be implemented according to different, rather arbitrary conventions. Moreover, because of the diversity of form for the various databanks, sharing of information could only be done on a case-by-case basis and with special programming. Use of TOD will reduce programming effort for users who store patient-related data.

2. Data Sharing

Because of the existence of the SCHEMA file, all the information required to allow sharing of data is in one place and in computer-readable form. This will allow data sharing between TOD users to occur more easily in the future.

3. High-Level Documentation

Aside from providing information to programs which operate on the TOD system, the SCHEMA file provides information to programmers and users

of a databank. This information, describing individual items in the databank, is defined by a Schema Language called DDL (Database Description Language) which uses a PL/ACME-like syntax for its declarations. This common language forms the basis for unambiguous communication among TOD databank groups. This communication process is strengthened by the fact that the different groups share a common general core set of programs and a common general file structure. Details of an individual databank are described using the Schema Language.

4. Operational Statistics

All the TOD programs store statistics which describe the operation of the databank. Careful review of these statistics in conjunction with the monthly summary of ACME charges will give the user a much clearer picture of what his computer dollar is buying.

5. Common Improvements

As ACME and users find ways to improve the TOD system programs and procedures in terms of capabilities and cost-effectiveness, these improvements will be passed along to TOD users by changes in the TOD programs and systems documentation to be implemented by means of monthly "releases" of the system. These releases will be upwards compatible. If a user writes a specialized program which he feels is worthy of sharing with the TOD group, this program can easily be generalized and made available as part of the TOD system.

Overview of the TOD System

The programs comprising the TOD system fall into four groups: data entry, data update, data retrieval and analysis, and TOD System Utility Programs. Figure I summarizes these groups.

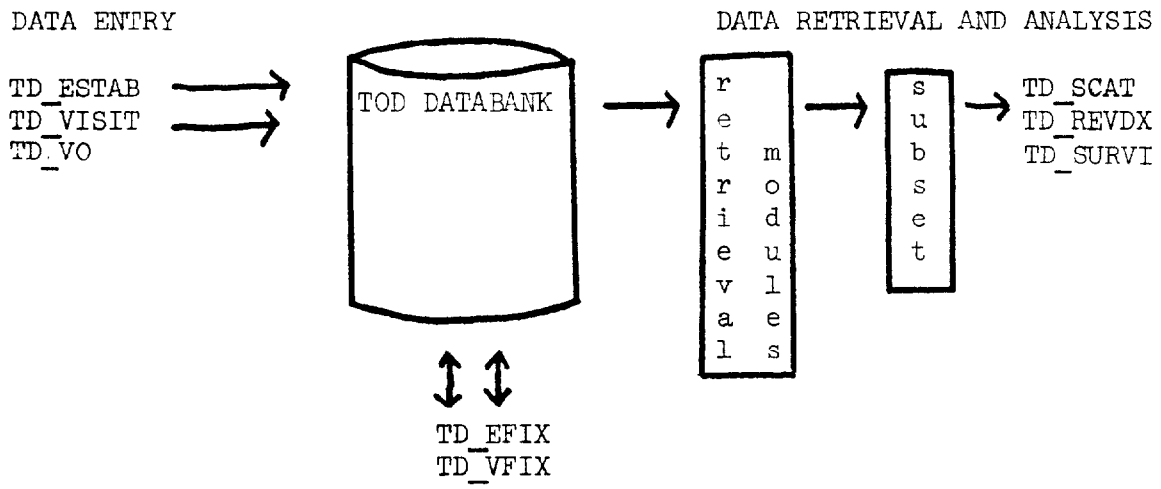


Figure I. TOD System Overview

TOD Programs

The programs named in Figure I are representative of the programs that will comprise the initial TOD system. Most of these are main programs, although a few are subprograms which can be included in a user-written program. Table I below summarizes the purpose of these programs.

Table I. TOD System Programs

Program	Class	Purpose
TD_ESTAB	entry	enter demographic (one-time) information associated with a patient.
TD_VISIT	entry	enter information measured at a point in time, e.g. at a patient clinic visit.
TD_EFIX	update	correct/change a patient demographic (one-time) data element.
TD_VFIX	update	correct/change information measured at a point in time.
TD_GCOL TD_GROW TD_GEXT TD_GROWX	retrieval sub-programs	routines to extract information from the TOD data files. These modules will be used by TOD programs and are available to the user for use in special purpose analysis programs.
TD_SALL TD_SAND TD_SOR TD_SSUPR	retrieval	create a library of subsets of patients and patient-visits based on the values of demographic, physiological, and general descriptive variables.
TD_GTDES	sub-program	extract schema information for a specific TOD databank.
TD_SUBST	sub-program	a routine used by all programs and available to extract those patients or patient-visits which satisfy certain criteria. The calling program then does its analysis on the reduced set.
TD_SCAT	analysis	construct scatter graph of two parameters
TD_REVDX	analysis	simple statistical review.
TD_QLIST	analysis	multi-optioned list databank contents program.
TD_SURVI	analysis	survival calculations. 4 methods. PUBLIC version of B. Brown's "Survival Kit".
TD_STENT	analysis	program to extract data from a TOD databank and write data file or pass array elements for the more common ACME statistical programs. (Not implemented.)

Program	Class	Purpose
TD_CHECK	utility	check data item values and file structure for inconsistencies.
TD_CSTCK	utility	operational statistics summary.
TD_TPOSE TD_RANGE TD_INDEX	utility	construct the indicated auxiliary file. Programs best run during times of low ACME utilization.
TD_DLIST	utility	Once the databank is defined, the schema is translated to a form more appropriate for computer processing.
TD_RECOM	utility	Periodically, users will wish to modify the form of their databanks by means of a major reorganization. This process, using an old and new schema (Database Definition), loads the new database from the old.

TOD Files

A TOD databank contains a number of inter-related files. Four of these files are required: td_schem, td_desc, td_head, and td_parm. In addition to these files, several auxiliary files can be added to the system to make retrieval of certain information faster. These files are td_index, td_range, td_htpse, and td_ptpse. Table II summarizes the purposes of these files.

Table II. TOD Files

File	Information Content
td_schem	Description of the databank in PL/ACME DECLARE-type statements.
td_desc	Internal form of the databank description.
td_head	Demographic patient information. One record per patient.
td_parm	Information measured at a point in time. One record per time per patient.
td_index	(HEADER item value, KEY to HEADER file) pairs sorted on header values. One such group for each header element that is indexed.
td_range	For each patient, the hi and lo ranged parameter values across all parameter records associated with the patient over time. Only those parameters which are ranged are included.
td_htpse td_ptpse	For TRANSPOSED data items these files contain the same information as the HEADER and PARAMETER files except that the ordering is such that all values for a particular item are contiguous, making questions which relate to specific items much faster to answer.

The Purpose of the Present TOD Effort

The present TOD implementation is not to be the system to end all information retrieval systems. Its capabilities have been limited in order to assure that a demonstratable working system can be swiftly implemented. Nevertheless, a full set of capabilities are provided to handle most of the users who are following patients over time. Once a number of TOD users exist, who speak a common language, further extensions to the system can be planned in a meaningful manner.

ACME views the TOD system as a set of programs which allows users who follow patients over time to set up, maintain, and use a databank in a simple and efficient manner. The present TOD effort is a study of the patient databank question in the Stanford Medical Center.

Further Reading

A reference to all ACME notes describing the operation and use of various portions of the TOD system and its implementation is given in ACME Note TODREF.

Revision of TODI-2 dated March 16, 1973.
Dist: Staff/TOD/All

ACME Note

Index to TOD ACME Notes

TODREF-1
Steve Weyl
April 4, 1973

This note is a comprehensive index to the set of ACME Notes describing the TOD (Time-Oriented Databank) system. The index is given in three parts: Part I references the notes that all planners and users should be familiar with. Part II references the file structure and system implementation notes, which are primarily of interest to systems analysts and programmers. Part III references historical notes, notes describing administrative procedures for the TOD system, and notes associated with individual TOD databanks.

The TOD system is a set of programs available from ACME designed to aid users in the creation, maintenance, and use of computer "databanks" which store patient-related information over time.

ACME notes TODI and TDOV give an overview of the TOD system. ACME note TODD is the original design document for TOD and is primarily of historical interest, since many of the conventions suggested there have been modified in the course of implementation.

* Notes marked with an asterisk (*) had not yet been published at the time this issue of TODREF went to press.

PART I - USE OF THE TOD SYSTEM

A. General Introduction and Overview

TODI — Introduction to the TOD (Time Oriented Databank) System -- F. Germano, S. Weyl

TDOV — TOD System Overview -- F. Germano

B. Planning and Defining a Databank

*TDPLAN — Planning a TOD-based Databank -- F. Germano, S. Weyl

*TDUA — How to Make a Schema for TOD -- V. Wiederhold

TODATA — Stanford Medical Center TOD Data Descriptor Dictionary -- F. Germano

TODDDL — The TOD Databank Description Language -- S. Weyl

TDPT — Definition of a TOD Databank Using PUBLIC Program TD_TRA -- S. Weyl

TDPDT — Detranslation of a Databank Schema Using PUBLIC Program TD_DTRA -- S. Weyl

TODPDN Obtaining a Proof Listing of the Schema File Using TD_DLIST -- F. Germano

TDPRE Redefinition of a TOD Databank Using TD_RECOM -- S. Weyl

C. Entering and Correcting Data

*TDUB How to Enter Data on TOD -- V. Wiederhold

TODPDG Checking Data Values and File Linkage Using Program TD_CHECK -- R. Giusti

D. Report Generating Programs

TODPDF Patient Chart Listing Program TD_PLIST -- R. Giusti

TODPDL Listing of TOD Header & Parameter Files Using TD_QLIST -- B. Bassett

TODPDN Obtaining a Proof Listing of the Schema File Using TD_DLIST -- F. Germano

E. Retrieval and Analysis Programs

TODPDD TOD Retrieval Module Summary Sheet -- F. Germano

TODPDO Definition of Patient Subsets for Analysis Using Programs TD_SALL, TD_SAND, TD_SOR, and TD_SSUPR -- S. Weyl

TODPDB TOD Scatterplot Program -- F. Germano

TODPDC TOD Reviewdx Program -- F. Germano

TODPDE TOD Survival Kit - User Instructions -- J. Whitner

TODPDJ TOD Debug Lister Program TD_QKLST -- R. Giusti

TODPDM Using TOD Retrieval Modules as Debug Programs -- R. Giusti

*TODSUR TOD Survival Kit - Computational Methods -- M. Hu

F. TOD Utility Programs

TODPDG Checking Data Values and File Linkage Using Program TD_CHECK -- S. Weyl, R. Giusti

TODPDH Construction of Range File Using TD_RANGE -- R. Giusti

TODPDI Construction of Transpose File Using TD_TPOSE -- S. Weyl, R. Giusti

TODPDK Constructing TOD Index Files with Program TD_INDEX -- R. Giusti

G. Writing Your Own Analysis Programs

TIDA TOD Analysis Programs -- F. Germano

H. Operational Costs of TOD Databanks

TODPDA Operational Overview for a TOD Databank -- F. Germano

TODCST Analyzing the Costs of Running a TOD Databank -- F. Germano

PART II - INTERNAL DOCUMENTATION

A. Program Documentation

TDSUB User-Supplied TOD Subprograms for Data Checking and Coding -- S. Weyl

TIDA TOD Analysis Programs -- F. Germano

TIDB TOD Operational Statistics -- F. Germano

TIDD Program PRE_PROC -- F. Germano

TIDF TOD Survival Kit - Structure and Linkage -- J. Whitner

B. File Structure

*TIDJ The TOD Data Files and Their Contents -- S. Weyl

TIDC The TRANSPOSE File -- F. Germano

TIDE Structure of the TOD Index File -- R. Giusti

TIDF TOD Survival Kit - Structure and Linkage -- J. Whitner

TIDG Record 1 in the TOD Descriptor File, td_desc -- F. Germano

TIDH Structure of the Subset Library File, td_subs -- S. Weyl

PART III - OTHER ACME NOTES

A. Historical

TODD Definition of the PL/ACME Time-Oriented Databank Protocol -- S. Weyl

DBT ACME Data Base for Cancer Virus Tumor Samples (Medical Microbiology - Dr. Hayflick) -- S. Weyl

DBD ACME Data Bases for Drs. Eugene Dong and Phillip Caves - Cardiovasular Surgery Research -- S. Weyl

MOP Comment on Medical Applications Oriented Preliminary Data Base -- S. Weyl

PMOD Need for a Medical Applications Oriented Data Base Protocol and Support Facility -- S. Weyl

BSPD Sharing Patient Data Files -- G. Wiederhold

DBS Present and Potential Patient-Related Databanks at the Stanford Medical Center -- F. Germano, G. Wiederhold

HTP Preliminary Data Base for Heart Transplant Pilot Research on Dogs -- S. Weyl

B. TOD Administrative Procedures

TODADM Administrative Procedures for the PL/ACME Time-Oriented Databank (TOD) -- F. Germano, S. Weyl

C. Notes on Individual TOD Databanks

TDUONA Programs PRELET - ONCOLET: Oncology Letter Writing Programs -- J. Whitner

*TDUONB Time-Oriented Databank for the Oncology Clinic -- S. Weyl

D. Keyword Index to TOD Notes

*TODIDX Keyword Index to TOD Notes -- F. Germano, S. Weyl

OTHER REFERENCES

1. Wiederhold, Gio, An Advanced Computer System for Medical Research, PROCEEDINGS OF THE IBM JAPAN COMPUTER SCIENCE SYMPOSIUM--Research and Development and Computer Systems

2. Frey, Girardi, Wiederhold, A Filing System for Medical Research, BIOMEDICAL COMPUTING, (2) (1971).

3. Wiederhold, Gio, Database Structures and Schemas (to be published)

4. Fries, James, Time Oriented Medical Research and a Computer Data Bank, JAMA, vol. 222, no. 12, Dec. 18, 1972, pp. 1536-1542.

Dist: Staff/TOD/All

Core Research & Development (Continued)

The TOD Databank Description Language

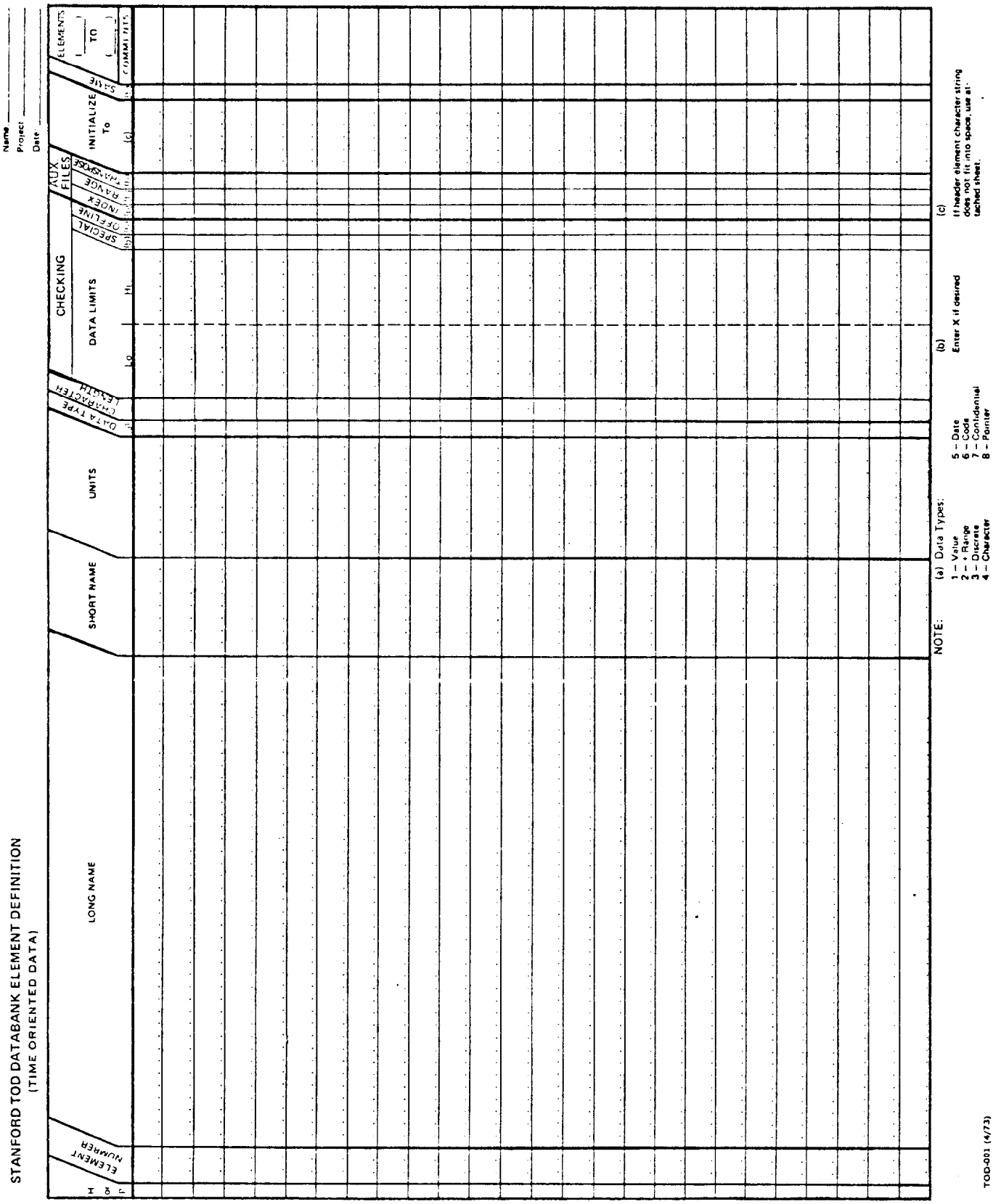
The TOD Databank Description Language is a means to define medical data in a less ambiguous form than has been used in the past. ACME Note TODDDL describes this defining capability.

In order to make the task of defining a patient databank easier, forms were designed which contained spaces for the same information required by the TOD Databank Description Language. A sample TOD Databank Element Definition form appears on the next page.

Once several databanks were defined using the TOD Databank Description Language, the concept of the TOD Data Descriptor Dictionary came into being. The Stanford Medical Center Data Descriptor Dictionary is a listing of the data elements in all the TOD databanks. This listing is arranged in order by the symbolic (short 8-character) name assigned to TOD data elements by individual databank planners. To each symbolic name a two-character suffix has been appended to indicate which TOD databank the data element resides in. When several databanks have the same symbolic name for a data item (which should only occur when the elements are indeed the same data variable in each of the individual databanks), they appear together in the listing, each with its own unique suffix.

The data dictionary pulls together in one place the variables stored in all the TOD databanks. It enables new databank planners to see what data already exists in other TOD databanks, but more importantly, it shows what conventions, such as data checking or units, were assigned to the data items.

A sample page for the TOD Data Descriptor Dictionary follows.



SAMPLE

TOD DATABANK ELEMENT DEFINITON FORM